THE AI TOOLCHAIN

NO. 017

Tail

THE DAILY RELEASE FIREHOSE

PUBLISHED SEPTEMBER 5, 2026 · EVERY WEEKDAY

The daily firehose — everything the toolchain shipped today, already filtered.

// HOW THIS ISSUE IS MADE

We read every release from the 187 tools on our watchlist at the source — **GitHub and GitLab release notes**, vendor **release pages** and **changelogs**, project **blogs and feeds**, vendor **press releases**, and the **source code** behind the tag. Bug-fix-only releases and non-product newsroom noise are dropped; what's left is summarized down to the new capability, how to try it, and any **screenshots or videos** the release itself published. Every entry links to the sources it was built from.

VIEW

ISSUE VIEW

full issue

FULL ISSUE | EXPAND IN PLACE | PREVIEW PANE

Do you prefer **full issue**? 👍 👎

▸ // CONTENTS SHOW 31 tools

▾ // THE BRIEF HIDE SAME ISSUE, SAME PROMPT, TWO WRITERS: GPT-5.5 Opus 5

A read across the whole issue before you read any of it: what stands out in today's releases, grouped by what it lets you do. Every tool named links to its entry below.

Three unrelated things stand out today: SGLang can benchmark scheduler and KV-cache choices without tying up GPUs; v0 lets teams fix preview-branch CI, merge, and deploy failures from chat; ai-gateway now carries request identity through provider attempts so routing failures are traceable end to end.

DEPLOY

Tune inference scheduling without burning accelerator time

SGLang added a no-GPU Simulator for benchmarking scheduling and KV-cache configurations. That makes capacity and latency experiments cheaper to run before touching production hardware, instead of discovering bad batching or cache settings under real load.

SGLang

BUILD

Move a generated app from chat to GitHub without leaving the workflow

v0 now publishes through isolated GitHub preview branches and lets users resolve CI, merge, and deployment issues directly from chat. The handoff from AI-generated changes to repository hygiene becomes less manual: failed checks and deployment blockers can be handled where the change was created.

v0

OPERATE

Trace an AI request across gateway retries and provider calls

ai-gateway now propagates request identity end to end and records per-attempt routing outcomes, while LangSmith added OpenTelemetry resource metadata and enforceable per-project trace limits. Operators get a clearer path from an app request to the gateway and upstream provider traces, with guardrails that stop one project from flooding trace storage.

ai-gateway · LangSmith

GOVERN

Turn agent and app security checks into repeatable evidence

GitHub Copilot CLI added a built-in security-review agent and enterprise content-exclusion enforcement; xalgorix expanded deterministic pentest verifiers to CORS, BFLA, BOLA, POST reflected XSS, and CSRF; agent-safe-pipeline can verify Decision Dossiers offline or against production endpoints with Ed25519 signatures. These reduce ad hoc review by

> making security findings and governance artifacts reproducible instead of screenshots and one-off prompts.
>
> GitHub Copilot CLI · xalgorix

Does **GPT-5.5** read better? 👍 👎

DEPTH BRIEFED **EVERYTHING** Every tool in the issue, in full.

OFFENSIVE SECURITY

◆ Exploitation & C2

xalgorix

5 RELEASES · 2026-09-04 NOTES ↗ LEAD

The xalgorix platform runs AI pentesting agents for reconnaissance, vulnerability detection, and exploitation workflows.

> xalgorix expanded its AI pentesting benchmark harness with new CORS, BFLA, and two-account BOLA challenges, extended verify_xss to confirm POST-based reflected XSS, and shipped verify_csrf — completing its deterministic verifier family covering SQLi, SSTI, XSS, XXE, OOB, and CSRF.

└──▷ WHAT SHIPPED · 5 FEATURES most completely described first ? what's the number?

01 **POST-based reflected XSS verification** IMPROVED 93

Extends `verify_xss` to confirm POST-based reflected XSS by passing the form action as `url`, a urlencoded form body as `data` (e.g. `search=<payload>`), with `method` defaulting to `POST` whenever `data` is present; the verifier stages a self-submitting cross-origin form so the payload executes exactly as it would for a victim.

Confirm a reflected XSS that is only reachable via a POST form submission — previously unverifiable without manual browser steps.

```
$ verify_xss(url="https://example.com/search", data="search=<payload>")
```

— Runnable example with exact parameters and mechanism v4.6.53

02 **Authenticated two-account BOLA challenges** NEW 85

Adds authenticated, two-account challenge support to the benchmark harness, wiring role A as the scan session and role B as the authorization comparison identity via `XALGORIX_TARGET_AUTH` / `XALGORIX_TARGET_AUTH_B`. Adds a `bola` challenge that serves every order to any logged-in user with no per-object ownership check, producing a CWE-639 IDOR/BOLA signal via cross-account differential (card_last4 PII leak), plus a `safe-bola` negative-control challenge that enforces per-object ownership (403) to verify a scanner does not over-report.

— Names env vars, CWE, and cross-account mechanism v4.6.58

03 **verify_csrf completes deterministic verifier family** NEW 85

Adds `verify_csrf`, a Cross-Site Request Forgery confirmer that replays a forged Origin / Referer request without an anti-CSRF token, reusing the scan session's

cookies, and confirms CSRF when the server accepts it (2xx/3xx with no token/forbidden rejection); records CWE-352 evidence and declines on `Authorization` `-header-protected` endpoints to avoid false positives on token-auth APIs. This completes the deterministic verifier family — `verify_sqli`, `verify_ssti`, `verify_xss`, `verify_xxe`, `verify_oob`, and now `verify_csrf` — giving every common injectable or forgeable vulnerability class a one-call exploit-proven evidence path.

— Names mechanism, CWE, and complete verifier family list `v4.6.52`

---

04 **CORS benchmark challenge and scorer** NEW 80

Adds a `cors-credentialed-reflect` benchmark challenge where `/api/account` reflects any request `Origin` into `Access-Control-Allow-Origin` with `Access-Control-Allow-Credentials: true`, exposing `email` and `api_token` to any attacker origin, plus a safe-cors negative control that reflects only a single fixed trusted origin. Adds a CORS text-signal check (matching 'CORS', 'cross-origin resource sharing', or 'access-control-allow-origin') evaluated before CWE lookup so generic access-control findings (e.g. `CWE-284`) aren't folded into the idor class, plus `CWE-942` / `CWE-346` -> `cors` canonical-class mappings and `cors` canonicalClass variants.

— Names endpoint, headers and CWE mappings; no runnable example `v4.6.64`

---

05 **BFLA benchmark challenge coverage** NEW 65

Adds a `bfla` challenge exposing an admin-only user-list function (emails + SSN fragments) that checks authentication but not caller role, enabling vertical privilege escalation testing scored under the `idor` class. Adds a `safe-bfla` negative control that enforces role checks (regular users receive 403) to catch over-reporting scanners.

— Describes challenge mechanism and scoring; no runnable example `v4.6.60`

WAS THIS USEFUL? 👍 👎

BUILD

◆ AI Coding Agents

Warp ⓘ 1 RELEASE · 2026-09-05 CHANGELOG ↗ ALSO

Warp is an AI terminal that runs commands and helps developers build software.

> Warp introduced Warp Factories, an open infrastructure for building software factories with an Activity view and YAML-based configuration, alongside smaller agent and CLI additions for retrieving harness conversation transcripts and overwriting files in one step.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 Conversation transcript retrieval for harness runs NEW

75

The `oz run get --conversation <run-id>` command retrieves the full conversation transcript from a third-party harness run, intended for post-incident review.

Retrieve the full conversation transcript from a third-party harness run for post-incident review.

```
$ oz run get --conversation <run-id>
```

— Exact runnable command given, minimal surrounding detail changelog-20260905-b04aafa8

02 Warp Factories software factory infrastructure NEW

70

Warp Factories provide "open, flexible infrastructure for building your software factory." An Activity view groups tasks into Triage, Planning, and Building, with linked issues, implementation plans, channels, and "Needs attention" statuses. Factories can be defined as "software factories as code" via a YAML configuration specifying repositories, agent models and roles, and a GitHub pull-request trigger.

— UI flow and config shape described, but no exact keys/commands given

changelog-20260905-b04aafa8

03 File overwrite flag for agent create_file action NEW

70

The `create_file` agent action now accepts an `allow_overwrite=true` parameter, letting an agent overwrite an existing file in a single step without a preceding delete or edit — useful when a code-gen agent needs to regenerate a file from scratch, e.g. create_file(path='src/config.py', content='...', allow_overwrite=true).

Overwrite an existing file in one agent step without a preceding delete or edit, useful when a code-gen agent needs to regenerate a file from scratch.

```
$ create_file(path='src/config.py', content='...',
allow_overwrite=true)
```

— Exact parameter and example call given, but scope beyond this is unclear

changelog-20260905-b04aafa8

WAS THIS USEFUL? 👍 👎

StackBlitz bolt.new

1 RELEASE · 2026-09-05 | CHANGELOG ↗ | ALSO

Bolt.new is an AI-powered web development platform that builds websites and applications from prompts.

> bolt.new shipped a reorderable prompt queue for stacking follow-up instructions mid-build, a unified project deletion flow that cleans up domains and databases in one step, and image preview inside Code view.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 **Prompt queue with reorder and pause** NEW — 72

Bolt now lets you queue up to 20 prompts to send while it is still building, preserving per-prompt agent selection. Queued prompts can be edited, reordered, deleted, or paused before Bolt processes them, and in shared projects each queued prompt shows the collaborator's avatar so teams can see who queued what.

— Describes mechanism and limits but no exact UI path or command

changelog-20260905-42932b16

02 **Unified project deletion and cleanup** NEW — 72

A single deletion action now disconnects custom domains, unpublishes the live site, deletes the Bolt database, and removes project files and chat history in one step. Before deleting, you can export the project to keep a local copy, and Bolt reports any resources it cannot clean up itself, such as external Supabase databases or third-party DNS records, so you know what to remove manually.

— Names concrete cleanup scope and manual-removal reporting, no exact menu path given

changelog-20260905-42932b16

THINNER COVERAGE BELOW

03 **Image preview in Code view** NEW — 40

Clicking an image file in Code view now shows an image preview.

— Clear action (click image file) but minimal detail beyond that

changelog-20260905-42932b16

WAS THIS USEFUL? 👍 👎

Vercel v0

1 RELEASE · 2026-09-05 | CHANGELOG ↗ | LEAD

The v0 generative UI platform builds web applications from natural-language prompts.

> v0 introduces a unified GitHub publishing workflow that isolates changes in preview branches and lets users resolve CI, merge, and deployment issues directly from chat.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 Unified branch-and-Publish workflow for GitHub projects NEW

85

A new Publish action for GitHub-backed projects creates or reuses a pull request, merges it into the base branch, and deploys to production in one step. Each change is auto-committed to an isolated working branch with its own preview deployment for validation before merging, and a branch menu lets users visit the latest preview deployment, review the code diff, create/open/merge the pull request, check CI checks and repository-rule status, and pull recent base-branch changes. Users can also ask v0 in chat to fix preview, CI, or merge issues without leaving the interface.

Validate a change in an isolated preview before it reaches production, then ship it without leaving v0.

> 📍 1. Make your changes in v0 chat — each change is auto-committed to an isolated working branch and a preview deployment URL appears in the branch menu. 2. Open the branch menu and select 'Visit the latest preview deployment' to validate the app. 3. When satisfied, select Publish — v0 creates or reuses the pull request, merges it into the base branch, and deploys the merged result to production.

Unblock a stuck deploy by asking v0 to diagnose and fix a failing CI check without switching to another tool.

> 📍 1. Open the branch menu and review CI checks and repository-rule status. 2. If a check is failing, type a prompt in the v0 chat such as 'Fix the failing CI check' — v0 analyzes the issue and applies a fix to the working branch. 3. A new preview deployment is created; re-check CI status in the branch menu before selecting Publish.

— Names the Publish action, branch menu options, and chat-based fixes with clear steps.

changelog-20260905-a8a1bc7e

WAS THIS USEFUL? 👍 👎

Replit Agent ⓘ 1 RELEASE · seen 2026-09-05 NOTES ↗ ALSO

Replit Agent is an AI coding agent that builds, tests, and deploys applications in Replit.

> Replit Agent adds daily scheduled database backups with configurable retention, plus a new Project Analytics dashboard for published web apps including Agent-driven funnel analysis on paid plans.

└──▷ **WHAT SHIPPED · 2 FEATURES** most completely described first ? what's the number?

01 **Project Analytics dashboard for published apps** NEW 75

Adds Project Analytics in the Growth pane showing visitor trends, popular pages, traffic sources, countries, browsers, and devices for published web apps, enabled by turning on Enable analytics and republishing. Also enables Agent-driven funnel analysis for actions such as sign-ups and purchases on paid plans via Project Analytics.

Start collecting visitor analytics for a published web app so you can see traffic sources and popular pages in the Growth pane.

> 📍 In your Replit project, open the Growth pane, toggle on 'Enable analytics', then republish your app.

— Names UI location and toggle, plus paid-plan funnel analysis detail. snapshot-20260905

02 **Scheduled daily production database backups** NEW 70

Adds scheduled backups that create one full restore point per day for your production database, with a configurable retention period set under Keep backups for in production database settings.

Enable daily database backups with a custom retention window so you can recover production data after an incident.

> 📍 In your Replit project, go to the production database settings, locate 'Keep backups for', set your desired retention period, then toggle on scheduled backups.

— Names exact setting and behavior, but no restore mechanics detailed. snapshot-20260905

WAS THIS USEFUL? 👍 👎

Daytona

1 RELEASE · seen 2026-09-05 NOTES ↗ ALSO

Daytona provisions isolated development sandboxes for AI coding agents through an API and SDK.

> Daytona 0.204.0 expands snapshot management to work by name, adds outbound proxy routing for sandbox traffic, and introduces typed git transport errors for programmatic handling.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 Snapshot lookup by name and source lineage filter IMPROVED

71

Snapshot `delete` and `activate` operations now accept either an ID or a human-readable name, expanding beyond ID-only operations. The snapshot list operation also gains a `sourceSandboxId` filter parameter for targeted queries of snapshots derived from a specific source sandbox.

List only snapshots derived from a specific source sandbox to audit lineage.

```
$ curl -X GET 'https://api.daytona.io/v1/snapshots?
sourceSandboxId=<sandboxId>' \
  -H 'Authorization: Bearer <token>'
```

Activate a snapshot by human-readable name instead of opaque ID, useful in scripted pipelines.

```
$ curl -X POST https://api.daytona.io/v1/snapshots/my-snapshot-
name/activate \
  -H 'Authorization: Bearer <token>'
```

— Names exact params and operations, runnable examples given `snapshot-20260905`

02 Outbound proxy routing on sandbox create NEW

70

Adds an `outboundProxyUrl` parameter to the sandbox create API call, allowing sandbox traffic to be routed through an outbound proxy at creation time.

Route all outbound sandbox traffic through a corporate proxy at sandbox creation time.

```
$ curl -X POST https://api.daytona.io/v1/sandboxes \
  -H 'Authorization: Bearer <token>' \
  -H 'Content-Type: application/json' \
  -d '{"outboundProxyUrl":
"http://proxy.corp.example.com:3128"}'
```

— Names exact param with runnable API call example `snapshot-20260905`

THINNER COVERAGE BELOW

03 **Typed git transport error codes in SDK** NEW 55

Introduces typed error codes `GIT_TRANSPORT_FAILED` and `GIT_REMOTE_REJECTED` in the SDK for programmatic handling of git remote failures.

— Names exact error codes but no usage example given `snapshot-20260905`

WAS THIS USEFUL? 👍 👎

Command Code

1 RELEASE · seen 2026-09-05 NOTES ↗ INDEX

The first AI coding agent that learns your coding taste. Powered by taste-1, a meta neuro-symbolic model.

> Command Code v1.49.0 adds support for GPT-6 Astra as a selectable model.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **GPT-6 Astra model support** NEW 35

Command Code adds GPT-6 Astra as a supported model, giving users a new model option for the coding agent.

— Names the model but no mechanism or setup steps. `v1.49.0`

WAS THIS USEFUL? 👍 👎

Graphify

1 RELEASE · 2026-09-05 NOTES ↗ ALSO

Graphify converts code, docs, SQL schemas, configs, and PDFs into a queryable knowledge graph with explainable relationships.

> Graphify's query scorer now ranks matches on a node's `rationale` attribute as a distinct tier, improving recall for rationale-only hits without diluting exact-match scores.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **Rationale attribute scoring tier in query scorer** IMPROVED

57

The query scorer ranks a match on a node's `rationale` attribute as its own scoring tier, placed between an exact label match and a source-path match. This adds recall for rationale-only hits without inflating exact-match coverage scores.

— Names exact scoring tier order but no usage example or command shown. v0.9.54

WAS THIS USEFUL? 👍 👎

Anthropic Claude Code ⓘ 1 RELEASE · 2026-09-04 NOTES ↗ CODE ↗

ALSO

Claude Code is Anthropic's terminal coding agent that plans, edits, and tests code in local repositories.

> Claude Code v2.1.261 adds command-line controls for subagent prompts and output size, a `/skill-doctor` command for pruning unused skills, org-policy diagnostics, and a batch of VS Code extension improvements spanning MCP server management, session list, model picker, and output styles.

└──▷ WHAT SHIPPED · 9 FEATURES most completely described first ? what's the number?

01 Inline output size limits for bash and tasks NEW 80

Adds `bashOutputMaxChars` and `taskOutputMaxChars` settings to control how much command and background-task output Claude receives inline before spilling to a file, configurable up to 128K characters.

Raise the inline output cap so Claude sees up to 128K characters of bash command output before it falls back to a file.

```json
{
  "bashOutputMaxChars": 131072,
  "taskOutputMaxChars": 131072
}
```

— Named config keys with exact limit and runnable JSON example. v2.1.261

02 Subagent system prompt from file NEW 75

Adds `--append-subagent-system-prompt-file` flag to read the subagent system prompt from a file, enabling prompts too large to pass on the command line.

Pass a large subagent system prompt from a file when it would be too long for the command line.

```
$ claude --append-subagent-system-prompt-file ./subagent-prompt.txt
```

— Exact flag with runnable command example, brief on mechanism. v2.1.261

03 Faster Vertex AI startup with service account credentials

IMPROVED

70

Improves startup on Google Vertex AI when `GOOGLE_APPLICATION_CREDENTIALS` is set: API client creation no longer re-runs project discovery or spawns extra `gcloud` processes.

— Names env var and exact mechanism fixed, but no benchmark given. v2.1.261

04 **/skill-doctor command for unused skills** NEW

60

Adds `/skill-doctor` slash command to show which loaded skills go unused and what they cost in context tokens, so you can prune them.

— Named slash command with clear purpose, no example shown. v2.1.261

05 **Model picker shows human-readable names** IMPROVED

60

Improves the `/model` picker and the VS Code model pill to show a human-readable model name instead of its raw Bedrock, Vertex AI, or LLM gateway ID, and changes the VS Code model picker to one flat list of every model, with rows for older model spellings listed last.

— Names the surfaces affected and the before/after display change. v2.1.261

THINNER COVERAGE BELOW

06 **Organization policy diagnostics** NEW

55

Adds an 'Organization policy' diagnostic line to `/status` and `claude doctor` explaining why an org policy could not be loaded (e.g., a proxy blocking the endpoint).

— Names both diagnostic surfaces but no example of output. v2.1.261

07 **Custom output style walkthrough** NEW

50

Adds a 'Build a custom style' walkthrough to the Output styles menu in VS Code that writes a custom output style file and lists it immediately.

— Names the menu and outcome, no file path or format given. v2.1.261

08 **MCP server management in VS Code extension** NEW

45

Adds an Add server form and a Remove action to the MCP servers dialog in the VS Code extension so MCP servers can be managed without leaving the IDE.

— Clear UI starting point, no mechanism detail beyond the dialog. v2.1.261

09 **Fold button for permission and question prompts** NEW

40

Adds a fold button to permission and question prompts in the VS Code extension so the conversation behind them can be read without dismissing them.

— Describes behavior but no named control or setting. v2.1.261

WAS THIS USEFUL? 👍 👎

OpenAI Codex CLI ⓘ 1 RELEASE · 2026-09-04 NOTES ↗ CODE ↗ INDEX

OpenAI Codex CLI runs an agent in the terminal that reads, changes, and tests code in local repositories.

> OpenAI Codex CLI added GPT-6-Astra to the Amazon Bedrock model picker, expanding model choice for Mantle and Runtime global/US routes.

└──▷ **WHAT SHIPPED · 1 FEATURE** most completely described first ? what's the number?

01 **GPT-6-Astra added to Bedrock model picker** NEW 50

GPT-6-Astra is now selectable in the Amazon Bedrock model picker for Mantle and Runtime global/US routes.

— Names model and routes but no mechanism or config detail `rust-v0.153.3`

WAS THIS USEFUL? 👍 👎

GitHub Copilot CLI ⓘ 1 RELEASE · 2026-09-04 NOTES ↗ LEAD

GitHub Copilot CLI is a terminal-based coding agent that can plan, write, and test code in local repositories.

> GitHub Copilot CLI adds a built-in security-review agent and enterprise content-exclusion enforcement, alongside a dense v1.0.83 update covering model fallback configuration, sandbox dev-tool and network controls (including breaking changes to local sandbox networking), MCP/CIMD OAuth, mTLS proxy support, and UI consolidation across sessions, sandbox, and plugins.

└──▷ **WHAT SHIPPED · 15 FEATURES** most completely described first ? what's the number?

01 Model selection and fallback configuration NEW 90

Adds a `model-policy: required` config key alongside a multi-value `model` list in custom agents, which is tried in order until an available model is found and is kept as the active list for subsequent turns; also adds support for the `claude-fable-5.1` model.

Lock a custom agent to a preferred model list so it never silently falls back to an unapproved model in a regulated environment.

yaml

```yaml
model:
  - claude-fable-5.1
  - claude-sonnet-4-5
model-policy: required
```

— Named config keys plus a runnable YAML config example v1.0.83

02 Sandbox network egress restrictions and Allow local network toggle

85

BREAKING

Sandboxed commands on macOS and Linux can no longer reach services running on the local machine; on macOS this also blocks servers the command itself starts on 127.0.0.1, so test suites that bind a local port will fail. Linux sandboxing now requires `slirp4netns`, `nsenter`, `iptables`, `ip6tables`, `iptables-restore`, and `ip6tables-restore` on PATH, and Linux sandbox network egress is now restricted to the configured proxy, which in proxy mode additionally requires `slirp4netns`, `util-linux` 2.35+, `iptables`, and `/dev/net/tun` access. Use the new `Allow local network` toggle in `/sandbox` to restore localhost access on macOS and Linux.

— Full mechanism and named binaries, with a clear remediation toggle v1.0.83

03 **Relative path resolution change for --add-dir and --plugin-dir** 80

BREAKING

A relative `--add-dir` or `--plugin-dir` path now resolves against the session's working directory under `--resume=<id>` and `--worktree`, and after `-C` is applied, rather than the directory the CLI was launched from — scripts that relied on the old resolution order will break.

— Exact flags and before/after behavior with migration implication v1.0.83

---

04 **Sandbox developer-tool path access control** NEW 80

Adds `sandbox.allowDevToolAccess` config key to control whether sandboxed file tools can read developer-tool paths such as `~/.npmrc`; set to `false` to revoke those grants.

Prevent sandboxed commands from inheriting npm token paths (e.g. ~/.npmrc) when running untrusted code.

yaml

```yaml
sandbox:
  allowDevToolAccess: false
```

— Named config key with concrete path example and runnable config v1.0.83

---

05 **Security-review specialist agent** NEW 75

Adds a built-in `security-review` specialist agent invokable with the `/security-review` slash command, which searches the codebase for exploitable vulnerabilities and reports only high-confidence findings.

Run a focused, high-confidence vulnerability scan on your codebase without switching tools.

```
$ /security-review
```

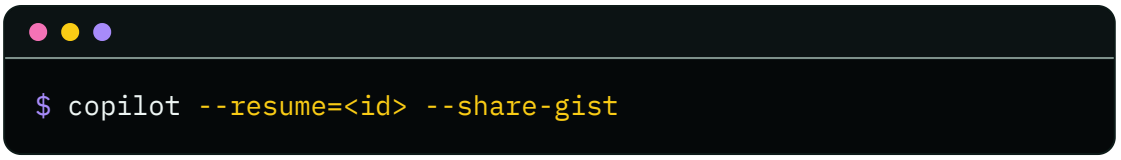
— Runnable slash command with clear scope described product docs

---

06 **Session transcript sharing flags** NEW 70

Adds `--share` and `--share-gist` flags to export a resumed session's whole transcript instead of only the latest run.

Export the full transcript of a long-running resumed session to a shareable gist for async review.



— Named flags with a runnable command example `v1.0.83`

07 **herdr terminal multiplexer support** NEW — 60

Adds `herdr` terminal multiplexer detection, enabling the Kitty keyboard protocol, color scheme following, terminal progress, `/copy`, and notifications in herdr panes.

— Names specific features enabled but no command to run `v1.0.83`

THINNER COVERAGE BELOW

08 **Plugin management UI consolidation** IMPROVED — 55

Moves `/mcp config` and the MCP add/edit/authenticate forms into the plugins dashboard instead of a separate MCP manager, and shows bundled built-in plugins in plugin list commands and `/plugin`.

— Named commands but only a UI navigation path `v1.0.83`

09 **Content exclusion policy enforcement** IMPROVED — 50

Copilot CLI now respects content exclusion policies configured at the enterprise, organization, and repository levels for Copilot Business and Copilot Enterprise users, ensuring excluded files are not used as context.

— Clear behavior change but no config surface named `product docs`

10 **Sessions sidebar sorting options** NEW — 50

Adds Recent, Created, Name, and classic None sorting to the split Sessions sidebar, with the selected order saved across restarts.

— UI feature with named options but only a navigation starting point `v1.0.83`

11 **Sandbox policy view improvements** IMPROVED — 45

Improves the `/sandbox` policy view by grouping path grants by source and showing detected developer tools.

— UI path given but no further mechanism `v1.0.83`

12 **Enterprise sign-in organization pinning** NEW — 40

Adds the `forceLoginOrgs` managed setting for enterprise admins to pin sign-in to approved GitHub organizations.

— Named setting but no config location or example given `v1.0.83`

13 **HTTPS proxy mTLS client certificate support** NEW — 35

Adds automatic HTTPS proxy mTLS client certificate support for model and web requests.

— Named capability with no configuration detail `v1.0.83`

14 **MCP OAuth via Client ID Metadata Document** NEW 35

Adds Client ID Metadata Document (CIMD) support for MCP OAuth sign-in.

— Named protocol but no usage steps provided `v1.0.83`

15 **Windows 11 taskbar session status** NEW 35

Shows running Copilot sessions in the Windows 11 taskbar with live hover status cards.

— Describes the feature but no interaction detail `v1.0.83`

└──▷ BREAKING ON UPGRADE

! On macOS and Linux, sandboxed commands can no longer reach services running on the local machine; on macOS this also blocks servers the command itself starts on 127.0.0.1 — test suites that bind a local port will fail. Use Allow local network in `/sandbox` to restore localhost access.

! Linux sandboxing now requires `slirp4netns`, `nsenter`, `iptables`, `ip6tables`, `iptables-restore`, and `ip6tables-restore` on PATH; sandboxed commands will fail to launch if these are missing.

! Linux sandbox network egress is now restricted to the configured proxy; proxy mode additionally requires `slirp4netns`, `util-linux` 2.35+, `iptables`, and `/dev/net/tun` access.

! A relative `--add-dir` or `--plugin-dir` path now resolves against the session's working directory under `--resume=<id>` and `--worktree`, and after `-C` is applied, rather than the directory the CLI was launched from — scripts that relied on the old resolution order will break.

WAS THIS USEFUL? 👍 👎

AGENT

◆ AI Agent Frameworks

PydanticAI

1 RELEASE · 2026-09-05 NOTES ↗ ALSO

PydanticAI is a framework that builds type-safe Python agents with dependency injection, model integrations, tools, and structured outputs.

> PydanticAI v2.40.0 focuses on realtime voice session improvements — mid-speech interruption handling, custom provider instantiation, and out-of-band prompt injection — alongside background price updates and a new agent event listener API.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 **Background model pricing updates** NEW 77

Adds pydantic_ai.prices.update_in_background() to keep model pricing data current without blocking the main thread, useful for long-running services that need accurate token-cost estimates.

Keep model pricing data refreshed in the background so token-cost estimates stay accurate in a long-running service.

```python
import pydantic_ai.prices

pydantic_ai.prices.update_in_background()
```

— Names exact function and purpose, runnable example given v2.40.0

02 **Agent event listener decorator** NEW 75

Adds an `@agent.on_event` decorator for registering event listeners directly on an `Agent` instance, letting code observe model responses, tool calls and other events without modifying the agent's tool or output logic.

Observe every agent event — model responses, tool calls, etc. — without modifying the agent's tool or output logic.

```python
from pydantic_ai import Agent

agent = Agent('openai:gpt-5.6-sol')

@agent.on_event
async def log_event(event):
    print('event:', event)
```

— Named decorator with runnable example, mechanism explained v2.40.0

03 **Realtime session control: barge-in, provider factory, send/enqueue**

NEW 72

Realtime sessions gain several controls: `handle_barge_in=True`, interrupt(played_bytes=...), and `played_audio_bytes` for mid-speech interruption handling; a `provider_factory` parameter on `infer_realtime_model` for custom provider instantiation logic; a `respond=` parameter on RealtimeSession.send() to control whether a text turn solicits a model reply; and RealtimeSession.enqueue() for injecting out-of-band prompts from code driving the session.

— All named surfaces listed but no usage example given v2.40.0

WAS THIS USEFUL? 👍 👎

LangChain

1 RELEASE · 2026-09-04 NOTES ↗ INDEX

LangChain is an open-source framework that orchestrates applications powered by language models.

> LangChain's langchain-core 1.6.2 release adds async tool support for OpenAI integrations.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 **Async tool support for OpenAI integrations** IMPROVED 28

`langchain-core` 1.6.2 adds async tool support for OpenAI integrations, letting tools used with OpenAI models be invoked asynchronously.

— No mechanism, API name, or usage example given beyond the bare statement.

`langchain-core==1.6.2`

WAS THIS USEFUL? 👍 👎

Agno (formerly Phidata) ⓘ 1 RELEASE · 2026-09-04 NOTES ↗

ALSO

Agno is an agent framework and runtime that orchestrates and runs multi-agent systems.

> Agno v3.0.6 focuses on MCP serving improvements — stateless transport, protocol-era negotiation, and a server discovery endpoint — alongside a new route-level auth exclusion, broader AgentOS upload support, and Bedrock client customization.

└──▷ WHAT SHIPPED · 6 FEATURES most completely described first ? what's the number?

01 **Stateless MCP serving mode** NEW 85

Adds MCPConfig(stateless=True) to serve `/mcp` without session tracking, enabling replica-agnostic routing in multi-instance deployments. This comes at the cost of server-initiated notifications and SSE resumability, and is off by default.

Deploy AgentOS MCP behind a load balancer with no sticky sessions — any replica handles any request.

python

```python
from agno.mcp import MCPConfig

mcp_config = MCPConfig(stateless=True)
# Pass mcp_config to your AgentOS app; all replicas can now
answer any /mcp request.
```

— Mechanism, tradeoffs, and runnable code example all given. v3.0.6

02 **MCP protocol era negotiation** NEW 85

Adds MCPTools(protocol_mode=...) to select the MCP protocol era to negotiate: `'legacy'` preserves session-based behaviour and is the default, while `'auto'` negotiates the newest era both sides support, using the sessionless protocol available from `2026-07-28`.

Negotiate the newest MCP protocol era automatically so your agent picks up sessionless transport once the server supports it.

```python
from agno.tools.mcp import MCPTools

tools = MCPTools(protocol_mode='auto')
# 'auto' will use the sessionless protocol (from 2026-07-28)
when both sides support it;
# falls back to legacy otherwise.
```

— Flag values, date, and runnable example fully specified. v3.0.6

03 **Public route exclusion in AuthorizationConfig** NEW 80

Adds `AuthorizationConfig.excluded_route_paths` to mark custom routes public without disabling auth globally, so specific paths can bypass JWT auth while all other routes keep it enforced.

Expose a custom route publicly while keeping JWT auth active on all other routes.

```python
from agno.auth import AuthorizationConfig

auth = AuthorizationConfig(
    excluded_route_paths=["/health", "/public/webhook"]
)
# All other routes continue to require a valid JWT.
```

— Config key demonstrated with a runnable, concrete example. v3.0.6

04 **MCP Server Card endpoint** NEW 65

Adds a `GET /mcp/server-card` endpoint that lists served tools and carries a configurable name, version, and instructions.

— Endpoint named with method and path but no usage example. v3.0.6

THINNER COVERAGE BELOW

05 **Zip and Eml file upload support in AgentOS** NEW 35

AgentOS now supports `.zip` and `.eml` file uploads.

— Formats named but no mechanism or example given. v3.0.6

06 **Custom async client for Bedrock model/embedder** IMPROVED

35

Supports passing a custom async client to the Bedrock model or embedder.

— Bare description with no example, code path, or scope detail. v3.0.6

WAS THIS USEFUL? 👍 👎

DEPLOY

◆ AI Model & Data Infrastructure

Perplexity API

1 RELEASE · 2026-09-05 | CHANGELOG ↗ | ALSO

Perplexity API Platform - build with the Router, Agent, Search, and Embeddings APIs. Real-time, web-wide research and Q&A capabilities for your products.

> Perplexity API added a new low-cost model, GLM 5.3 Flash, to its chat completions endpoint.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 **GLM 5.3 Flash model added** NEW 75

The `perplexity/glm-5.3-flash` model is now available via the Perplexity chat completions endpoint at $0.15/M input tokens, callable by setting `model` to `perplexity/glm-5.3-flash` in requests to `https://api.perplexity.ai/chat/completions`.

Call the new GLM 5.3 Flash model for cost-efficient inference via the Perplexity chat completions endpoint.

```
$ curl https://api.perplexity.ai/chat/completions \
  -H 'Authorization: Bearer <YOUR_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"model": "perplexity/glm-5.3-flash", "messages":
[{"role": "user", "content": "Summarize recent CVEs in
OpenSSL."}]}'
```

— Names exact model id, price, and runnable curl example `changelog-20260905-9d2f804d`

WAS THIS USEFUL? 👍 👎

Fireworks AI

1 RELEASE · 2026-09-05 | CHANGELOG ↗ | ALSO

Fireworks AI provides hosted inference, model fine-tuning, and deployment APIs for open-weight models.

> Fireworks AI expanded its Serverless LoRA training pool with three new base models and introduced tiered rate limits by model size, while deprecating several older models from both Serverless inference and Serverless Training.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 Serverless and Serverless Training model deprecations DEPRECATED

85

`GPT OSS 20B`, `GPT OSS 120B`, `Kimi K2.6 Turbo/Fast`, `Kimi K2.7 Code Fast`, `DeepSeek V4 Pro`, and DeepSeek V4 Pro (0813) are deprecated from Serverless effective August 27, 2026, after which workloads using these models will break. Separately, `Qwen 3.5 9B` and `Qwen 3.6 27B` are deprecated from Serverless Training effective August 26, 2026, and existing training workloads on these models must migrate to `Qwen 3.8 27B`.

— Names all deprecated models, exact dates, and migration target.

changelog-20260905-11636b9d

02 New base models for Serverless LoRA training NEW

60

The shared Serverless Training pool now supports LoRA workloads on `DeepSeek V4 Flash 0731` (up to 262K context), `Qwen 3.8 27B` (up to 128K context), and `Muse Glimmer 30B` (up to 128K context) as available base models.

— Names exact models and context limits but no usage steps. changelog-20260905-11636b9d

THINNER COVERAGE BELOW

03 Model-size-tiered serverless rate limits IMPROVED

55

Serverless adaptive rate limit ceilings now vary by model size tier: models under 400B parameters receive higher ceilings, models between 400B and under 1.6T receive intermediate ceilings, and models at or above 1.6T retain the previous base ceilings.

— Explains tier mechanism and thresholds but no config surface.

changelog-20260905-11636b9d

└──▷ BREAKING ON UPGRADE

! GPT OSS 20B, GPT OSS 120B, Kimi K2.6 Turbo/Fast, Kimi K2.7 Code Fast, DeepSeek V4 Pro, and DeepSeek V4 Pro (0813) are deprecated from Serverless effective August 27, 2026 — workloads using these models will break after that date.

! Qwen 3.5 9B and Qwen 3.6 27B are deprecated from Serverless Training effective August 26, 2026 — existing training workloads using these models must migrate to Qwen 3.8 27B.

WAS THIS USEFUL? 👍 👎

Groq

1 RELEASE · seen 2026-09-05 NOTES ↗ LEAD

Groq is a high-speed inference engine that runs large language models significantly faster than traditional GPUs.

> Groq's only shipment this window was the addition of OpenAI's open-weight GPT-OSS models to its inference API, bringing built-in browser search and code execution at very high token throughput.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 **OpenAI GPT-OSS 20B and 120B models added** NEW 95

Groq added `openai/gpt-oss-20b` and `openai/gpt-oss-120b` as new models served via `POST /openai/v1/chat/completions`. Both offer a 131K token context window, 32K max output tokens, built-in browser search and code execution, and structured output support; the 20B model runs at ~1000+ TPS while the 120B model runs at ~500+ TPS.

Run a reasoning query against the fast 20B model to triage time-sensitive security incidents at ~1000+ TPS.

```
$ curl https://api.groq.com/openai/v1/chat/completions \
  -H "Authorization: Bearer $GROQ_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{"model": "openai/gpt-oss-20b", "messages": [{"role":
"user", "content": "Analyze this log and identify signs of
lateral movement: <log>"}]}'
```

Use the 120B model for deep multilingual threat-intelligence analysis where accuracy outweighs throughput.

```
$ curl https://api.groq.com/openai/v1/chat/completions \
  -H "Authorization: Bearer $GROQ_API_KEY" \
  -H "Content-Type: application/json" \
  -d '{"model": "openai/gpt-oss-120b", "messages": [{"role":
"user", "content": "Summarize this threat report and extract
IOCs"}]}'
```

— Full specs, endpoint and runnable curl examples given for both models. snapshot-20260905

WAS THIS USEFUL? 👍 👎

HeyGen HyperFrames ⓘ 1 RELEASE · 2026-09-05 NOTES ↗ ALSO

Write HTML. Render video.

> HyperFrames v0.8.28 adds frame-diffing for CI regression testing, a new lint check for mismatched media sources, and resolution scaling for transparent renders.

└──▷ **WHAT SHIPPED · 3 FEATURES** most completely described first ? what's the number?

01 Snapshot frame-diffing against reference video NEW 80

The CLI `snapshot` command gains an `--against` flag that pairs each rendered frame against a reference video, enabling regression diffing in CI pipelines.

Catch frame-level regressions in CI by diffing the current render against a known-good reference video.

```
$ npx hyperframes snapshot --against reference.mp4
composition.html
```

— Named flag with runnable example command showing exact usage. v0.8.28

THINNER COVERAGE BELOW

02 Lint error for mismatched media-kind sources NEW 50

HyperFrames now raises a lint error when a `<video>` or `<img>` src points to the wrong media kind, such as an image URL used as a video source.

— Describes trigger condition but no command to invoke lint. v0.8.28

03 Resolution scaling for alpha output IMPROVED 25

Adds resolution scaling support for alpha (transparent) output renders.

— Bare mention with no mechanism or usage detail. v0.8.28

WAS THIS USEFUL? 👍 👎

◆ Local LLM Runtimes

llama.cpp ⓘ

1 RELEASE · 2026-09-04 | NOTES ↗ | CODE ↗ | LEAD

LLM inference in C/C++

> llama.cpp's v0.4.0 release adds lazy tensor loading to cut startup memory, expands multimodal pipelines with video/audio input and finer-grained tokenization, adds sparse flash attention and an Apple RDMA RPC transport for distributed inference, ships support for five new model architectures, and reworks KV-cache handling in a way that breaks existing saved sessions.

└──▷ WHAT SHIPPED · 11 FEATURES most completely described first ? what's the number?

01 **Video and audio input for multimodal pipeline** NEW 85

Adds `--video-*` CLI flags and `mtmd_helper_init_opt` server options for video input in multimodal pipelines; the server now accepts `data:` URLs for `input_video` and `input_audio` fields in multimodal requests, and a new `mtmd_input_part` struct plus mtmd_tokenize_from_parts() API enable fine-grained multimodal tokenization.

Send a video clip as a data: URL to a multimodal server endpoint — no file hosting required.

```
$ curl http://localhost:8080/v1/chat/completions -H 'Content-
Type: application/json' -d '{"messages":
[{"role":"user","content":[{"type":"text","text":"Describe this
video"},
{"type":"input_video","input_video":"data:video/mp4;base64,
<base64data>"}]}]}'
```

— Names every flag, struct and API added, with runnable request example. v0.4.0

02 **Lazy tensor loading for model startup** NEW 75

Adds `--lazy-mode` CLI flag enabling on-demand (lazy) tensor reading to reduce startup memory overhead, plus `llama_lazy_mode` and `lazy_mode` in the C API for programmatic control.

Load a large model without pre-reading all tensors into RAM — useful on memory-constrained hosts where full model load would OOM.

```
$ llama serve --lazy-mode -hf ggml-org/Qwen3.5-0.8B-GGUF
```

— Clear flag and API names, runnable example, but no perf numbers. v0.4.0

03 **KV cache reuse and restore fidelity, breaking session format**

65

BREAKING

Adds n-gram history lookup via a KV-cell sequence position index to improve KV cache reuse, and adds KV-cell token tracking to improve context restore fidelity; this bumps session/state versions, so existing saved sessions will not be loadable after upgrading.

— Explains mechanism and flags breaking migration impact, but no config key. v0.4.0

---

04 **Per-layer CPU FFN offload for MoE models** NEW

65

Adds a per-layer CPU FFN offload capability for mixture-of-experts models too large to fit entirely on GPU VRAM, exposed via the `--n-cpu-ffn` flag.

Offload FFN layers to CPU for MoE models too large to fit entirely on GPU VRAM, using the new per-layer CPU FFN flag.

```
$ llama serve -m deepseek-v4.gguf --n-cpu-ffn 8
```

— Runnable example names the flag, though no bullet describes internals. v0.4.0

---

05 **Sparse flash attention for DeepSeek-V4/GLM and Qwen4exp** NEW

60

Adds sparse flash attention via `ggml_flash_attn_ext_set_n_kv_max` for DeepSeek-V4/GLM and Qwen4exp models.

— Named function and target models, no usage example. v0.4.0

THINNER COVERAGE BELOW

06 **New model architecture support** NEW

55

Adds initial support for Qwen3.8-Flash-Next (`qwen4exp`), NVIDIA Nemotron-3-Puzzle-75B-A9B, DSpark for Nemotron 3.5, nanbeige4.2-3B, and the DeepSeek-V4-Flash-Vision-Exp multimodal model.

— Names five models but no usage instructions. v0.4.0

---

07 **RPC transport and async execution for distributed inference** NEW

50

Adds Apple RDMA as an RPC transport backend for distributed inference, and adds RPC event/async backend APIs enabling asynchronous execution and allocation-dependency tracking.

— Names both additions but lacks further API detail. v0.4.0

---

08 **Quantizer RAM cap via max_buf_size** NEW

45

Adds `max_buf_size` to quantize params, capping quantizer RAM usage during large-model quantization.

— Named param but no usage example or values given. v0.4.0

---

09 **Preserve chain-of-thought reasoning by default** IMPROVED

40

Enables `preserve_reasoning` by default in the server, preserving chain-of-thought content in responses.

— Named config flag but no example or override syntax. v0.4.0

---

10 **Per-slot context limit in server** NEW

35

Adds a per-slot context limit to the server, allowing each inference slot to have its own context cap.

— Describes capability but no config key or flag named. v0.4.0

---

11 **Autoscaled YaRN training context** IMPROVED

30

Autoscales YaRN training context when YaRN rope scaling is specified, removing the need for manual tuning.

— States behaviour change but no parameters or values. v0.4.0

└──▷ BREAKING ON UPGRADE

! Session and state file versions are bumped for KV-cell token tracking; existing saved sessions will not be loadable after upgrade.

WAS THIS USEFUL? 👍 👎

vMLX

1 RELEASE · 2026-09-04 | NOTES ↗ | ALSO

The vMLX server runs compressed MLX models on Apple Silicon with disk caching, paged memory, continuous batching, and hybrid SSM scheduling.

> vMLX 1.6.53 is a focused hardening release for MTP speculative decoding, adding a reversible AR safety valve, tunable native-MTP verify parameters, proposal-head improvements for Qwen models, and a finalized Flash-Next MTP bundle for qwen4_exp.

└──▷ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 Proposal-head enhancements for MTP (Qwen3.5/3.8 VLM, Qwen4) NEW

65

Adds a q4 proposal head for the Qwen3.5/3.8 VLM MTP lane (default off), a one-time per-bundle proposal-head stamp honored on later launches to avoid repeated head evaluation at startup, and enables the `qwen4_exp` q4 proposal head ON by default for Qwen4 models.

— Names default states and model targets, still no explicit flag v1.6.53

02 Flash-Next MTP bundle finalization for qwen4_exp NEW

65

Ships the final Flash-Next MTP bundle contract for `qwen4_exp` (index-driven weights plus calibrated sidecar), and adds a one-time dated redownload warning in the panel for the Flash-Next MTP fix, localized and conformance-aware.

— Names bundle contract and warning surface, points to panel UI v1.6.53

THINNER COVERAGE BELOW

03 Reversible AR safety valve for MTP NEW

58

Adds a policy-independent, on-the-fly AR safety valve for MTP that is windowed and context-scaled, running in both text and multimodal lanes. It pairs with a reversible AR fallback that tracks a measured AR tier and uses MTP re-entry probes so the engine can re-enter speculative decoding after a demotion.

— Describes mechanism and lanes but no flag or command to invoke it v1.6.53

04 Native-MTP verify tuning: depth ceiling and projection padding NEW

50

Adds a parameterized native-MTP depth ceiling for measurement, now covering verifier and profile caps, alongside stock-MLX projection padding for native-MTP verify (default off).

— Names the tuning surfaces but no exact key or command `v1.6.53`

WAS THIS USEFUL? 👍 👎

DATA

◆ Vector DBs & RAG

ai-memory

1 RELEASE · 2026-09-04 | NOTES ↗ | ALSO

Solution for long term memory for agent coding CLIs and to facilitate handoff between different agent vendors

> ai-memory 2.0.3 widens `finalize-session --agent` to cover all captured agent types and adds filesystem free-space visibility to status and migration output.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 Broader agent support in finalize-session

IMPROVED

70

`finalize-session --agent` now accepts all captured agents: `hermes`, `claude-desktop`, `crush`, and `other`, letting sessions from these agents be finalized without a separately registered agent profile.

Finalize a session captured by Claude Desktop or Hermes without needing a separately registered agent profile

```
$ ai-memory finalize-session --agent claude-desktop
ai-memory finalize-session --agent hermes
```

— Names exact flag and all accepted agent values with a runnable example v2.0.3

THINNER COVERAGE BELOW

02 Free space shown in status and migration log

IMPROVED

45

`ai-memory status` output and the pre-migration log now surface filesystem free space.

— Names the command but no example or detail on thresholds v2.0.3

WAS THIS USEFUL? 👍 👎

Pinecone

1 RELEASE · 2026-08-19 | NOTES ↗ | INDEX

Pinecone is the vector database for AI agents and applications, built for semantic search, knowledge retrieval, and long-term memory at scale.

> Pinecone introduced Nexus BYOC, a new bring-your-own-cloud deployment model, and doubled the pod-to-serverless migration limit to 50 million records.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 **Higher pod-to-serverless migration limit** IMPROVED 50

The record limit for migrating a pod-based index to serverless has been raised to 50 million records, up from 25 million.

— Clear before/after numbers but no command shown `snapshot-20260905`

02 **Nexus BYOC deployment model** NEW 40

Pinecone introduces Nexus BYOC (Bring Your Own Cloud), a new deployment model with its own architecture distinct from Database BYOC, documented separately to clarify the relationship between the two.

— Names the model but not concrete setup steps or mechanism `product docs`

WAS THIS USEFUL? 👍 👎

EVALUATE

◆ AI Observability & Evals

LangChain LangSmith

1 RELEASE · 2026-08-10 NOTES ↗ LEAD

LangSmith is a platform for debugging, testing, and monitoring LLM applications built with LangChain.

> LangSmith's latest release adds enforceable per-project trace limits, a thread evaluator validation endpoint, OpenTelemetry resource metadata support, and bulk dataset split management in experiments, alongside breaking changes to bulk export compression, dataset comparison APIs, and ingestion log format.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 OpenTelemetry resource attribute metadata on traces NEW

80

LangSmith traces can now carry OpenTelemetry resource attributes as `otel.resource.*` metadata without modifying span emission code, populated via the standard `OTEL_RESOURCE_ATTRIBUTES` environment variable, e.g. `user.id=user-42,deployment.environment=production`.

Attach OpenTelemetry resource attributes (e.g. user ID, environment) to LangSmith traces as `otel.resource.*` metadata without modifying span emission code.

```
$ export OTEL_RESOURCE_ATTRIBUTES='user.id=user-
42,deployment.environment=production'
```

— Exact env var and metadata field named, ready to set `snapshot-20260905`

02 Batched-run ingestion log format change BREAKING

60

The batched-run ingestion log now emits `run_verbs` as a list of `run_id` and `verbs` objects instead of a map keyed by run UUID, which may break structured-log aggregators that expected the previous map format.

— Exact field name and shape change, no migration tooling given `snapshot-20260905`

└──▷ BREAKING ON UPGRADE

! Legacy dataset comparison helpers are removed from the public OpenAPI spec and generated SDKs; existing HTTP routes continue to work only for LangSmith UI clients.

! Bulk export compression default changes from gzip to zstandard (`zstd`); self-hosted deployments must set `FF_BULK_EXPORT_DEFAULT_COMPRESSION` to retain gzip.

! The batched-run ingestion log now emits `run_verbs` as a list of `run_id` and `verbs` objects instead of a map keyed by run UUID, which may break structured-log aggregators that

expected the previous map format.

---

ai-gateway

1 RELEASE · 2026-09-05 NOTES ↗ LEAD

Unified AI Gateway for 30+ LLMs (OpenAI, Anthropic, Bedrock, Azure etc) with Caching, Guardrails, A/B test & cost controls. Go-native Fastest & Scalable AI Gateway LiteLLM & Kong AI Gateway alternative.

> ai-gateway v1.5.4 adds end-to-end request identity propagation and deeper distributed tracing, letting operators see per-attempt routing outcomes and tie gateway traces to upstream provider traces.

└──▷ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 **Per-attempt routing trace spans** NEW 92

Adds `observability.tracing.attempt_spans` config key (default `false`) to emit one `gateway.routing.attempt` `CLIENT` child span per routing attempt — retries and failovers included — carrying `ferro.routing.target_key`, `ferro.routing.sequence`, and `ferro.routing.outcome`; backends plug in via the `observability.AttemptSpanProvider` interface.

Enable per-attempt child spans to see exactly which targets were tried, in what order, and with what outcome in your distributed trace.

yaml

```yaml
observability:
  tracing:
    attempt_spans: true
```

— Names config key, span attributes and provider interface with runnable example. v1.5.4

02 **Upstream trace propagation on pass-through routes** NEW 90

Adds `observability.tracing.propagate_passthrough` config key (default `true`) to forward W3C `traceparent` and `tracestate` headers upstream on `/v1/*` pass-through and fixed-target routes (`/v1/responses`, `/v1/files`, `/v1/batches`), joining the provider's trace to the gateway trace.

Disable upstream trace context injection on pass-through routes when you do not want providers joining the gateway trace.

yaml

```yaml
observability:
  tracing:
    propagate_passthrough: false
```

— Exact config key, default and affected routes given with example. v1.5.4

03 **Request identity propagation across spans, logs and events** NEW

90

Adds request identity propagation via the OpenAI `user` field, `X-User-ID` and `X-Session-ID` request headers, or W3C `baggage` entries `user.id` / `session.id`; recorded as `enduser.id`, `session.id`, and `ferro.request.metadata.<key>` on the request span, as `User`, `SessionID`, and `Metadata` on `observability.RequestAttrs`, `Event`, and `RoutingAttempt`, and as `user_id` / `session_id` on request-log rows (`GET /admin/logs`, schema migration 6). Embedders can attach request metadata programmatically via `observability.ContextWithRequestIdentity`.

— Names all input surfaces, storage fields and log endpoint, no runnable example. v1.5.4

04 **Metadata field added to RoutingAttempt and RequestAttrs types**

85

BREAKING

`observability.RoutingAttempt` and `observability.RequestAttrs` each gain a `Metadata` map field: unkeyed composite literals no longer compile (switch to keyed fields), neither type is usable as a map key, and `==` comparison is unavailable — use `reflect.DeepEqual` instead.

— Gives exact migration steps and affected types. v1.5.4

└──▷ **BREAKING ON UPGRADE**

! `observability.RoutingAttempt` and `observability.RequestAttrs` each gain a `Metadata` map field: unkeyed composite literals no longer compile (switch to keyed fields), neither type is usable as a map key, and `==` comparison is unavailable — use `reflect.DeepEqual` instead.

WAS THIS USEFUL? 👍 👎

Arize Phoenix

1 RELEASE · 2026-09-04 NOTES ↗ INDEX

Arize Phoenix is an open-source AI observability platform for tracing and evaluating LLM applications.

> Phoenix v20.8.0 adds filtering capabilities to the trace list endpoint and introduces MiniMax as a supported LLM provider.

└──▷ WHAT SHIPPED · 2 FEATURES most completely described first ? what's the number?

01 **Error-status and latency filters on trace list endpoint** NEW

65

The trace list server endpoint now supports `error-status` and `latency` filters, enabling programmatic filtering of traces by error state and latency.

— Names endpoint and exact filter params but no usage example arize-phoenix-v20.8.0

THINNER COVERAGE BELOW

02 **MiniMax LLM provider integration** NEW

35

Phoenix adds MiniMax as a supported LLM provider integration.

— Just names the provider, no mechanism or setup detail arize-phoenix-v20.8.0

WAS THIS USEFUL? 👍 👎

Langfuse

2 RELEASES · 2026-09-04 NOTES ↗ INDEX

Open source AI engineering platform: LLM evals, observability, metrics, prompt management, playground, datasets. Integrates with OpenTelemetry, LangChain, OpenAI SDK, LiteLLM, and more. YC W23

> Langfuse overhauled the Experiments UI around score-based comparison and rebuilt result views, while also tightening public-API auth enforcement across all migration modes.

└──▷ WHAT SHIPPED · 3 FEATURES most completely described first ? what's the number?

01 Experiments UI overhaul: scoring, comparison, and layout

IMPROVED 75

The Experiments results grid was rebuilt around score columns, adding a `worse than` comparison filter and a score-by-run matrix view for spotting underperforming runs. The chart grid was replaced with a compact metric strip, the empty Analytics tab was dropped, table settings were merged into a single unified panel, and the comparison view is now selectable, grouped, and pre-picked by default.

Compare experiment runs and identify underperformers using the new worse-than-comparison filter and score-by-run matrix in the Experiments UI.

> 📍 In the Langfuse UI, go to Experiments › select an experiment › open the Results grid › use the 'Worse than comparison' filter to surface runs scoring below a baseline, and switch to the score-by-run matrix view to compare scores across all runs at a glance.

— Names UI elements and flow but no config or API surface v4.30.0 v4.29.0

THINNER COVERAGE BELOW

02 Public API auth policy enforcement via policy seam IMPROVED

45

Public-API project routes are now gated through the policy seam in all migration modes, enforcing auth checks consistently across the API surface.

— No endpoint names or migration mode details given v4.29.0

03 Filter analytics tracking in evaluations NEW

23

Adds filter analytics tracking to the evaluations (evals) feature.

— Bare one-line addition with no mechanism described v4.29.0

WAS THIS USEFUL? 👍 👎

Braintrust ⓘ 1 RELEASE · 2026-09-01 NOTES ↗ ALSO

Ship quality agents at scale. Braintrust is the AI observability platform for tracing production, running evals, and catching regressions before they reach users.

> Braintrust's latest releases add scheduled analysis and automation around its Loop agent (Patterns and Loop automations), a new built-in multimodal model, time-window-based score alerting, and a set of SDK and eval-action refinements including blind human reviews and Harbor plugin output capture.

└──▷ **WHAT SHIPPED · 8 FEATURES** most completely described first ? what's the number?

01 PR comment score/metric filtering in eval action NEW

83

Adds `report_scores` and `report_metrics` inputs to the GitHub eval action (`v2.1.0`) to filter which scores and metrics appear in the PR comment, each accepting a comma- or newline-separated list of names.

Filter a PR comment to show only the scores and metrics you care about, reducing noise in high-signal CI pipelines.

yaml

```yaml
- uses: braintrust-dev/eval-action@v2.1.0
  with:
    api_key: ${{ secrets.BRAINTRUST_API_KEY }}
    report_scores: accuracy, faithfulness
    report_metrics: latency, cost
```

— Exact input names, version, and a runnable workflow snippet `snapshot-20260905`

02 GLM-5.3 Flash built-in model NEW

75

Adds `glm-5.3-flash` as a built-in multimodal reasoning model available via the Braintrust provider in playgrounds, prompts, and scorers, or via the Braintrust Gateway, with no AI provider setup required.

Use GLM-5.3 Flash as a zero-setup multimodal scorer in the Braintrust Gateway without configuring a third-party AI provider.

```
$ curl https://api.braintrust.dev/v1/gateway/chat/completions \
  -H 'Authorization: Bearer <BRAINTRUST_API_KEY>' \
  -H 'Content-Type: application/json' \
  -d '{"model": "glm-5.3-flash", "messages": [{"role": "user",
"content": "Score this response for accuracy: ..."}]}'
```

– Runnable Gateway curl example makes this immediately usable snapshot-20260905

03 **Blind human reviews setting** NEW 65

Adds blind human reviews: a project setting ('Blind human reviews') that hides peer scores, comments, and aggregates until a reviewer submits their own scores, with reviewers holding the project Update permission always exempt.

– Names the setting and exempt permission, no exact navigation path snapshot-20260905

04 **Agno eval instrumentation in Python SDK** NEW 63

Adds Agno eval instrumentation to the Python SDK (`v0.36.0`) — `AccuracyEval` , `AgentAsJudgeEval` , `ReliabilityEval` , `PerformanceEval` , and eval suites — traced automatically via auto_instrument().

– Names all eval classes and the entry function but no example snapshot-20260905

05 **Harbor plugin verifier output upload** NEW 63

Adds standard verifier output upload to the Harbor plugin in the Python SDK (`v0.36.0`): `test-stdout.txt` , `test-stderr.txt` , and `ctrf.json` , with `attachments` and `redact_patterns` configuration.

– Names output files and config keys but no usage example snapshot-20260905

THINNER COVERAGE BELOW

06 **Patterns scheduled failure analysis** NEW 50

Adds Patterns, a scheduled analysis feature that runs Loop against your project to surface recurring failure modes and cost trends, saving each finding with supporting traces and a suggested fix.

– Describes mechanism but no config key or command shown snapshot-20260905

07 **Scheduled Loop automations** NEW 50

Adds Loop automations: scheduled Loop runs with their own instruction, model, and write permissions, with results deliverable to a Slack channel or webhook.

– Explains behaviour but no exact setup surface named snapshot-20260905

08 **Time window alerts for score degradation** NEW 50

Adds Time window alerts to detect and alert on LLM score degradations that persist over a configurable time window, with support for webhook-based external aggregation as an alternative.

– Names the alert type but no config key or UI path given product docs

WAS THIS USEFUL? 👍 👎

GOVERN

◆ AI/LLM Security

Doberman-Core

1 RELEASE · 2026-09-05 NOTES ↗ ALSO

Your AI's guard dog. Doberman sits at runtime, gating every input, output and tool call to stop unsafe or unintended actions before they execute.

> Doberman-Core v0.18.6 rebuilds its setup flow with dry-run previews, turns the dashboard and TUI into full decision browsers, ships an experimental Cursor native-hooks adapter, and makes plugin activation explicit after upgrades — a breaking change for anyone relying on auto-loaded custom rules or detectors.

└──▷ WHAT SHIPPED · 5 FEATURES most completely described first ? what's the number?

01 Explicit plugin activation after upgrade

BREAKING

80

Plugins — custom rules, detectors, and auth-provider plugins — no longer auto-load on upgrade. Each must be explicitly reactivated with `doberman plugins enable <name>`.

Re-enable a custom detection plugin after upgrading, since plugins no longer auto-load in v0.18.6.

```
$ doberman plugins enable <name>
```

— Exact command and named plugin types, but no migration tooling described v0.18.6

02 Rebuilt `doberman setup` with dry-run preview

IMPROVED

65

`doberman setup` is rebuilt with a dry-run preview, honest exit codes, and a closing doctor pass.

— Named command and three concrete changes, no example run shown v0.18.6

THINNER COVERAGE BELOW

03 Dashboard and TUI as decision browsers

IMPROVED

50

The dashboard and TUI become real decision browsers, gaining filters, keyboard shortcuts, and a full explanation view.

— Describes new capabilities but no navigation path or command given v0.18.6

04 Experimental Cursor native-hooks adapter

NEW

50

Adds an experimental native-hooks adapter for Cursor with a session heartbeat and a doctor check.

— Names the integration and two mechanisms but no setup command v0.18.6

05 **Simplified auth dialog flow** IMPROVED 20

The auth dialog now asks one question per screen.

— Bare description with no mechanism or scope detail v0.18.6

└──▷ BREAKING ON UPGRADE

! Plugins (custom rules, detectors, auth-provider plugins) no longer auto-load on upgrade — each must be re-enabled with `doberman plugins enable <name>`.

WAS THIS USEFUL? 👍 👎

ai-safe2-framework

1 RELEASE · 2026-09-04 NOTES ↗ ALSO

The Universal Governance, Risk, Compliance (GRC) Operating System with Integrated Security for Agentic AI, Non-Human Identities, and Swarm Governance. AI SAFE² + AI Sovereignty Maturity Model (AISM), NEXUS-A2A Protocol, FORGE-Act, Marshal Plan for AI [Dual License: MIT + CC-BY-SA]

> AI SAFE² v3.1 ships a unified `safe2` command-line interface, structured AISM Decision Cards for translating assessments into governance decisions, new evidence-source integrations for NEXUS and NVIDIA SkillSpector, and a 27-requirement Unbiased AI Standard (UAS) regulatory profile extension.

└──▷ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 **Unified `safe2` command-line interface** NEW 80

Adds the `safe2` CLI with subcommands including `safe2 scan project` and `safe2 example verify` for scanning AI projects, evaluating evidence, and validating packaged examples in CI/CD and automated agent workflows.

– Names exact runnable subcommands and their CI/CD use

2026-09-04_Agent_CLI_AISM_Decision_Suppo…

02 **Unbiased AI Standard (UAS) regulatory profile** NEW 63

Adds the Unbiased AI Standard (UAS) as a regulatory profile extension with 27 requirements — 14 system, 8 human, and 5 bridge — without altering the core 161-control structure or creating CP.11.

– Precise requirement counts and structural constraints given, but no usage steps

2026-09-04_Agent_CLI_AISM_Decision_Suppo…

THINNER COVERAGE BELOW

03 **AISM Decision Cards for assessment findings** NEW 55

Adds AISM Decision Cards that transform technical assessment findings into structured summaries covering scores, evidence quality, conflicts, alternatives, pros/cons, success/failure ranges, decision owner, review date, and exit criteria.

– Lists card contents in detail but no invocation path given

2026-09-04_Agent_CLI_AISM_Decision_Suppo…

04 **NEXUS and SkillSpector evidence integrations** NEW 48

Adds NEXUS integration to incorporate NEXUS findings as structured evidence within AI SAFE² assessments, connecting implementation observations to governance and assurance requirements, and adds NVIDIA SkillSpector as an evidence source for

assessing agent skills and implementation risks, retaining source attribution and evidence context.

— Names both integrations but no config or endpoint detail

2026-09-04_Agent_CLI_AISM_Decision_Suppo…

WAS THIS USEFUL? 👍 👎

AI MODELS

◆ Frontier Models

Google Gemini API ⓘ 1 RELEASE · 2026-09-05 CHANGELOG ↗ ALSO

Build with Gemini 2.0 Flash, 2.5 Pro, and Gemma using the Gemini API and Google AI Studio.

> Gemini API shipped GA availability for gemini-3.8-flash, introduced the Lyria 3.5 music generation model in public preview, and added agentic video understanding to several Flash models for far more token-efficient long-form video analysis.

└──▷ WHAT SHIPPED · 1 FEATURE most completely described first ? what's the number?

01 **GA release of gemini-3.8-flash** NEW 60

`gemini-3.8-flash` reached general availability and is callable via the GenerateContent API (e.g. `models/gemini-3.8-flash:generateContent`) for tasks such as agentic software engineering and complex enterprise workflows.

Call the new `gemini-3.8-flash` model for an agentic software engineering or complex enterprise workflow task via the GenerateContent API.

```
$ curl -X POST
'https://generativelanguage.googleapis.com/v1beta/models/gemini-
3.8-flash:generateContent?key=$GEMINI_API_KEY' -H 'Content-Type:
application/json' -d '{"contents": [{"parts": [{"text": "Analyze
the following codebase and propose a refactoring plan."}]}]}'
```

— Named model and endpoint but no mechanism beyond GA status changelog-20260905-db92a6fe

WAS THIS USEFUL? 👍 👎

Anthropic ⓘ 1 RELEASE · 2026-09-03 NOTES ↗ ALSO

Anthropic is an AI safety company that develops Claude, a large language model assistant for text generation and analysis.

> Anthropic tightened validity and prefix-binding rules for extended thinking blocks in the Claude API, added mid-conversation tool changes and per-turn effort configuration, and shipped a new `ant apply` subcommand for declarative, file-based agent deployment.

└──▷ WHAT SHIPPED · 4 FEATURES most completely described first ? what's the number?

01 Thinking block validity checks and prefix-binding enforcement

95

BREAKING

The Claude API now validates every `thinking` block's origin model — for example Claude Fable 5.1 can read blocks from Claude Opus 5 but Claude Opus 5 cannot read blocks from Claude Fable 5.1, and unreadable blocks are silently dropped without billing. A new `block_binding.prefix_mismatch_behavior` field controls whether the API returns a 400 error or silently drops invalid thinking blocks when the request prefix changes, and responses now include an array listing any thinking blocks dropped during a request. For accounts created on or after August 31, 2026, 00:00 UTC, prefix-binding is enforced by default, rejecting or dropping blocks whose preceding context has changed.

Opt in to strict prefix-mismatch enforcement today (on older accounts) so your multi-turn thinking integration is ready before it becomes mandatory.

```json
{
  "model": "claude-fable-5-1",
  "thinking": {
    "type": "enabled",
    "budget_tokens": 10000,
    "block_binding": {
      "prefix_mismatch_behavior": "error"
    }
  },
  "messages": [
    {"role": "user", "content": "What is the GCD of 1071 and 462?"}
  ]
}
```

— runnable JSON config example with exact field name and enforcement date `product docs`

02 **ant apply for declarative agent deployment** NEW 71

Adds an `ant apply` subcommand to the `ant` CLI (v1.30.0) for declarative, file-based creation and updating of agents, environments, skills, memory stores, and deployments from repository files. Writes a `claude-lock.json` lockfile after each run so subsequent runs, locally or in CI, update existing resources instead of creating duplicates.

— named subcommand and lockfile, but no runnable example given `September 3, 2026`

THINNER COVERAGE BELOW

03 **Mid-conversation tool changes via tool_addition/tool_removal blocks**

NEW 48

Adds `tool_addition` and `tool_removal` block types that let tools be changed mid-conversation, replacing the previous need to edit earlier messages and invalidate thinking blocks.

— named block types but no usage example `product docs`

04 **Per-message effort settings via output_config** NEW 40

Adds per-message `output_config` support to change effort settings per turn without needing to edit prior messages.

— brief description, no example or field detail `product docs`

└──▷ BREAKING ON UPGRADE

! For accounts created on or after August 31, 2026, 00:00 UTC, the API enforces prefix-binding on thinking blocks by default — requests where prior messages have been edited will be rejected with a 400 error or have blocks dropped, depending on `block_binding.prefix_mismatch_behavior`.

WAS THIS USEFUL? 👍 👎

// END OF TRANSMISSION

The AI Toolchain · curated daily · September 5, 2026
Releases & features from the open-source and commercial AI tools that practitioners actually run.

▷ Subscribe